DOCUMENT RESUME

ED 204 359 TM 810 335

AUTHOR Shoemaker, Sharon H.; Johnson, Richard T.
TITLE Assessing the Construct Validity of a Criterion-Referenced Test: A Nomological Network Approach.
PUB DATE Apr 81
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Reseach Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; Educational Objectives; Elementary School Mathematics; *Evaluation Methods; Grade 6; Intelligence Quotient; Intermediate Grades; Norm Referenced Tests; Scaling; Sex Differences; Socioeconomic Status; *Test Validity; Time Factors (Learning)
IDENTIFIERS Guttman Scales; *Nomological Network; Variance (Statistical)

ABSTRACT

The applicability of a recognized construct validation procedure, the nomological network, to a sixth-grade criterion-referenced mathematics test developed in a metropolitan school system was investigated. Two samples were tested. The first sample was used to investigate two hypotheses: items within each objective will constitute a Guttman scale, and a principal-factor analysis will reduce the test items to groupings similar to the groupings on objectives set up by the test developers. Both hypotheses were supported. Results from the second sample indicated: (1) total test scores were higher on posttest than on pretest as a consequence of instruction on the objectives; (2) mean scores on objectives which received the most time-on-task showed the greatest gain when teachers knew the pretest objectives; (3) gain differences between males and females were not significant; (4) most criterion-referenced test scores were correlated with norm-referenced test scores on similar objectives; and (5) posttest variance was greater than pretest variance, contrary to the original hypothesis. Correlations of socioeconomic status and IQ with test scores were no greater for norm-referenced tests than for criterion-referenced tests. (BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
S.A. Shoemaker & R.T. Johnson
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED204359

TM 810 335

ASSESSING THE CONSTRUCT VALIDITY OF A CRITERION-REFERENCED TEST: A NOMOLOGICAL NETWORK APPROACH

by

Sharon H. Shoemaker and Richard T. Johnson
Northwest Regional Education Lab

Background

The need for evidence of the construct validity of criterion-referenced score interpretations has been strenuously emphasized (Fitzpatrick, 1980; Hambleton, Swaminathan, Algina, & Coulson, 1978; Messick, 1975). The purpose of this study was to investigate the applicability of a recognized construct validation procedure, the nomological network (Cronbach & Meehl, 1955; Grandy, 1979), to a sixth-grade criterion-referenced mathematics test developed in a metropolitan school system. Where outcomes of the procedure supported experts' interpretations of what criterion-referenced scores should represent, the test was considered to have construct validity. Discriminant evidence, whereby the measure was shown not to be unduly related to other variables, was also assessed (Messick 1975; Grandy, 1979).

Glaser introduced the concept of criterion-referenced testing (CRT) in 1963. During the subsequent 18 years, authors have classified and defined CRTs (Gray, 1978; Nitko, 1980) and have discussed the development, utilization, and validation of CRTs (Berk, 1978a, Popham, 1975; Roudabush & Green, 1972). But guidelines for establishing the construct validity of CRT scores have been lacking

Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, April, 1981.

(Hambleton et al. 1978; Linn, 1979). Messick (1975) stated that "all measurement should be construct-referenced" (p. 957), and "construct validity thus appears to be just as important for educational measurement as for psychological measurement, for there are no adequate options" (p. 961). Tenopyr (1977) concurred, "it is clear that processes and underlying constructs are involved in all measurement" (p. 48), and "content-oriented test development. . .must always take constructs into account" (p. 53).

Construct

The construct subject to interpretation was capability to perform mathematical tasks, or more simply, performance, as represented by specific item, objective, domain and total test scores. Although "capability to perform" is not identified as such in the literature, "performance" does reflect the thema running through the works of a number of authors (Berk, 1978b; Cox, 1971; Cronbach & Meehl, 1955; Garvin, 1971; Popham, 1971). This construct was derived from criterion-referenced test literature, as were components of the network. The network properties which were identified included: score sensitivity to classroom instruction; relationship between classroom time spent on individual objectives or domains and score change over the school year; hierarchical scalability of items within objectives; existence of specific objectives-factors; concordance between CRT domains and norm-referenced test (NRT) subscores on similar tasks; changes in test score variability after instruction; similarity of mean scores for males and females; and a lack of score relationship to both family socio-economic status (SES) based on the occupation of

the primary wage-earner in the home, and to non-verbal IQ as measured by the Ravens's Progressive Matrices Test.

Criterion-Referenced Test

A criterion-referenced achievement test of 100 multiple-choice items was selected for study. The test, designed for the sixth grade level, is a mathematics test covering 14 domains encompassing items representative of 32 individual objectives. In the development of the CRT, 200 items initially were piloted in open format. The best distractors along with the correct responses were incorporated into the items in a closed format and tested again. Data from these pilot testings were analyzed, and 100 items were selected for the final form, with the number of items per objective varying from one to nine, based on the degree of importance attached to the objective by the curriculum planners. The administration of these tests in May, 1980, marked the second year the tests were given to approximately 10,000 students.

Sample

Two samples of sixth-grade students were selected. Since the intent of this study was to demonstrate the applicability of a method of construct validity to CRTs, the similarity of the school district to others in the country is not critical. Only if the particular system was in some way exceptional and this peculiarity precluded method generalizability would a problem arise; other than a slightly higher mean SES than the U.S. as a whole, the district could be considered typical.

The first sample of 1,000 students was used to investigate two hypotheses:

(1) items within each objective will constitute a Guttman scale such that a student's passing a particular item on an objective will give information about his success on earlier ones for that objective; and (2) a principal-factor analysis will reduce the test items to groupings similar to the groupings on objectives set up by the test developers. Both hypotheses were supported.

Scalability

The scalability of the items within each objective was one network property fairly well supported by the results. The items should be scalable if the test and CRT theory are congruent, and results from Guttman scaling were used as evidence. A coefficient of reproducibility (Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975, p. 533) of at least .90 was accepted as supporting the hypothesis. Twenty-three of the 32 CRT objectives attained this level, and an additional three were single item objectives where scability was meaningless.

When Guttman scaling was applied at the domain level, rather than at the objective level, the results were less satisfactory. The coefficients of reproducibility met the required .90 level in only three of the 14 domains. Such results suggest that the further removed one gets from the objectives level, the less useful the CRT results become for instructional purposes.

Objectives and Factors

The product-moment correlation matrix based upon 0-1 scored item responses was factor analyzed to determine the principal-factor solution using the SPSS (Nie et al. 1975) PA2 option. Instead of obtaining only those factors with eigenvalues larger than one, as is typically done, factoring continued until 32 factors were obtained to correspond to the number of objectives in the CRT, and a varimax rotation performed. The results are shown in Table 1, with only non-zero loadings (based on a significance testing procedure in Harman, 1967, p. 163) presented.

---

insert table 1 here

The a priori classification by objective was compared with the empirical factor structure and the differences noted. To the extent that items constituting an objective had sizeable loadings on one factor, and one factor only, construct validity was demonstrated. For example, CRT items 22, 23 and 24 are from the objective "divide with fractions having unlike denominators." When analyzed, these items all loaded on a single factor. Inspection of Table 1 shows that a fairly good correspondence between factors and objectives exists.

The second sample of 109 students was selected to investigate four hypotheses, each discussed separately below.

(1) total test scores will be higher on posttest than on pretest as a consequence of instruction on objectives underlying the test. The hypothesis was supported; data were analyzed by repeated-measures ANOVA with sex of student and classroom as factors.

Score Sensitivity to Instruction

Five classrooms were used, with the individual student considered the unit of analysis. Correlations of pretest with posttest for the 107 students with complete data ranged from .52 to .85. The mean age of the total sample was 12 years, 1 month, with a standard deviation of 4.7 months.

---

insert table 2 here

The ANOVA (Table 2) shows that the difference between pretest and posttest means was far beyond what could be expected by chance under the null hypothesis. The overall pretest mean was 50.73 and posttest was 72.70 with standard deviations 14.82 and 16.55 respectively. The most parsimonious explanation for the difference is that it is the result of direct instruction on objectives underlying the test.

The significant F-value for between-classrooms indicated that the overall CRT means differed among the five. In contrast, whether the students were male or female apparently made no difference in overall mean scores, nor did student sex within a particular class.

The interaction of time with class, with an F of 6.89, was also significant--not only was there a difference between pretest and posttest but there was a significant difference between the individual classes in their change from one time to the next. No other interaction terms appeared to be other than zero: males and females changed at the same rate within each of the classes.

(2) Mean scores on objectives which are ranked by teachers as having received the most time-on-task in their classroom

instruction will show the greatest gain from pretest to posttest. Correlations of teacher time-on-task rankings with standardized mean gain scores were significantly greater than zero.

Time-on-Task vs. Change

In the five classes sampled, three of the teachers had information on pretest objectives, and they ranked their time-on-task by objective. Rank order correlations between standardized gain on the 32 objectives and time-on-task were .571, .658, and .328; all three values exceeded the value needed for rejection of the null hypothesis that rho = 0 at the .05 level.

The remaining two teachers ranked time-on-task on the 14 domains, a much less specific classification. When rankings were compared to standardized gain on domains, the correlations were .363 and .156; neither value exceeded the .456 required to reject the null hypothesis at the .05 level. Thus the general hypothesis was supported in three of the five classrooms, where the teachers directed their instruction toward the objectives rather than the domains, and time-on-task was compared to objectives.

(3) Any differences in scores between examinees who differ in sex, socio-economic status, non-verbal IQ, and norm-referenced test performance will be in accordance with expectations derived from the literature.

(a) differences in posttest performance between males and females will be less than .3 s.d.,

Male-Female Score Differences

Although the sample mean for the females was higher than the sample mean of the males, 74.13 compared to 69.48 on the posttest, the difference of 4.65 was not significant when compared to the standard error of the difference (3.25); i.e., the difference between the means was small compared to the variability within cells. The observed difference in the statistic turned out to be about .28 of a standard deviation. The lack of significance made any further test of the male-female difference of no consequence; there is no reason to believe that the population means are different for males and females on the posttest total CRT score.

(b) CRT scores will be highly correlated with norm-referenced test scores on similar objectives, especially on the pretest,

CRT-NRT Similarity

The actual items on the CRT and the SRA were examined in order to match the domains on the CRT with the NRT skill areas. There were ten identified skills tested on the latter compared to 14 on the former. As the CRT domains were much more specific and homogeneous than the NRT skill areas, in four instances two separate CRT domains had items coinciding closely with a single SRA skill area. One SRA skill area had no CRT domain corresponding to it and one CRT domain had no SRA skill area to match. The intercorrelation matrix for these matches may be found in Table 3 with correspondences delineated by heavy lines.

---

insert table 3 here

The correlations on many of the skill-domain pairings are relatively high. This appears to indicate that considerable overlap exists in the skills called for in the two tests, thus supporting this facet of the construct validity network. Yet it is risky, even though items appear to be similar, to attempt to equate some of the domains with the skill areas. Most of the intercorrelations are substantial, so that all the separate domains are correlated with all the skill areas. Specification of the particular pairings was based on perusing the scanty definitions in the SRA manual coupled with examining the items. Other investigators could conceivably develop another set of pairings which made the same conceptual sense but gave slightly different results. In addition, the SRA, on many items, called for reading ability while the CRT was constructed to minimize the influence of reading ability on capacity to solve the questions. With these constraints in mind, an examination of the intercorrelation matrix shows that all the pairings have r's that differ significantly from zero, which supports the first part of the subhypothesis. The last part, which suggests that the CRT-NRT correlation will be higher for the CRT pretest than for the CRT posttest, had no support whatsoever. The apparent statistical significance of one single difference out of 14 comparisons could be dismissed as occurring by chance.

Again, a possible implication of these results is that once one leaves the level of objectives specificity on CRTs, the CRT begins to resemble an NRT.

(c)correlations between IQ and CRT score, and between SES and CRT score will be higher on the pretest than on

the posttest, and

(d) correlations of IQ and SES with NRT scores will be higher than correlations of IQ and SES with CRT posttest scores.

CRT Relationship to IQ & SES

The hypothesis that pretest CRT scores will be more highly correlated with Raven's IQ and with SES than posttest CRT scores failed to be supported. The correlation between IQ and CRT pretest was .495 and between IQ and CRT posttest was .547. Based on the N of 74 students with complete data, and taking into account the r of .777 between CRT pretest and posttest, the difference between the two correlations failed to reach significance (.05 level). The correlations themselves are large enough that the population rho can be assumed to be greater than zero at a confidence level of .01 or more.

CRT Relationship to SES

For the relationship between socio-economic status and CRT pretest, the correlation was .245; between SES and CRT posttest the correlation was .200. Again the difference between the two correlations, based on a sample of 97, could not be assumed to be other than zero.

CRT vs. NRT in Relationship to IQ & SES

The covariation between CRT and Raven's scores had been anticipated, as had the relationship between the CRT and SES. But the

hypothesis considered here stated that the extent of the relationship was higher for the NRT than the CRT. Such was not the case.

The Raven's IQ score and posttest CRT correlated .547, a value not significantly different from the IQ and NRT correlation of .427. The one-tailed hypothesis was in the opposite direction from the obtained statistics, so no additional test was conducted.

Almost the same situation occurred in regard to socio-economic status. The CRT-SES correlation of .200 did not differ significantly from the NRT-SES one of .085. Again the statistics were in the opposite direction from that hypothesized; once more the obtained *t* of -1.62 with 95 *df* did not exceed the critical value at the .05 level.

A considerable overlap existed between the CRT and NRT scores, as evidenced by the *r* of .751 between them. The extent of this relationship, coupled with the similarity of the strength across the two types of tests, indicated that the CRT and NRT are assessing much the same thing. The network property that the CRT is distinctive in its comparative freedom from contamination by intelligence or socio-economic status was completely without support. In fact, from the statistics reported here for total score, this particular criterion-referenced test is just another achievement test.

(4) Test score variability will be greater on the pretest than on the posttest; quite the opposite occurred.

Variability Changes

The pretest variance was 219.51 and posttest variance 274.00, which with a pre-post correlation of .777 resulted in a *t*-value of -1.72, non-significant at the 5 percent level.

The logic of those writing in the CRT field appears reasonable. Both developmental and learning theory support the tenet that concomitant with increasing age and learning comes an increase in variability of scores. Inspection of statistics for commonly used achievement tests shows larger standard deviations with each successive grade. While in the ideal situation one might wish to take students at different levels of knowledge in a content area to some specified plateau after instruction, such a case may not exist in reality. Perhaps a more likely scenario is for students at some common no-knowledge base to move to quite different final levels. Or perhaps both conditions occur simultaneously, leading to no difference in variability from pretest to posttest, as appears to be the case in this study. Had a ceiling effect been present, as may occur with other CRTs, perhaps the hypothesis would have been supported.

Conclusion

Methods used in the study have general applicability to CRT construct validation. One major conclusion was that many of the expectations postulated in the literature for CRTs may be valid at the objectives level, but not at the total test level.

The determination of construct validity is a complex, multi-faceted process. The construction of a nomological network accommodates the multiplicity of elements in a construct validation study and brings order to the process. The intent of this study was to present an approach to construct validity of a criterion-referenced test--to pull together, in a cohesive fashion, a myriad of assertions made about CRT scores. The criterion-referenced mathematics test

facilitated the investigation and drew attention to possible weak areas of the test itself as well as questionable aspects of CRT theory.

REFERENCES CITED

Berk, R. A. Criterion-referenced test item analysis and validation. Paper presented at the First Annual Johns Hopkins University National Symposium on Educational Research, Washington D. C., October 1978a.

Berk, R. A. A consumers' guide to criterion-referenced test item statistics. NCME measurement in education, 1978b, 9.

Cox, R. C. Evaluative aspects of criterion-referenced measures. In W. J. Popham (Ed.), Criterion-referenced measurement: an introduction. Englewood Cliffs, N.J.: Educational Technology, 1971.

Cronbach, L. J. & Meehl, P. E. Construct validity in psychological tests. Psychological Bulletin, 1955, 54, 281-302.

Fitzpatrick, A. R. Construct and decision validation methods for criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980.

Garvin, A. D. The applicability of criterion-referenced measurement by content area and level. In W. J. Popham (Ed.), Criterion-referenced measurement: an introduction. Englewood Cliffs, New Jersey: Educational Technology, 1971.

Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.

Grandy, J. E. An investigation into the construct validity of three career skills exercises. Princeton, New Jersey: Educational Testing Service, 1979, 5, RDR79-80, No. 1.

Gray, W. M. A comparison of Piagetian theory and criterion- referenced measurement. Review of Educational Research, 1978, 48, 223-249.

Hambleton, R. K., Swaminathan, H., Algina, J. & Coulson, D. B. Criterion-referenced testing and measurement: a review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Harman, H. F. Modern factor analysis (2nd ed.). Chicago: University of Chicago Press, 1967.

Linn, R. L. Issues of validity in measurement for competency-based programs. In M. A. Bunda & J. R. Sanders (Eds.), Practices and problems in competency-based measurement. Washington D.C.: National Council on Measurement in Education, 1979.

Messick, S. The standard problem: meaning and values in measurement and evaluation. American Psychologist, October 1975, pp. 955-966.

Nie, N., Hull, C., Jenkins, J., Steinbrenner, K., & Bent, D. Statistical package for the social sciences (2nd ed.). New York: McGraw-Hill, 1975.

Nitko, A. J. Individual differences between criterion- referenced tests. Paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980.

Popham, W. J. (Ed.) Criterion-referenced measurement. Englewood Cliffs, New Jersey: Educational Technology Publications, 1971.

Popham, W. J. Educational evaluation. Englewood Cliffs, New Jersey: Prentice-Hall, 1975.

Roudabush, G. E. & Green, D. R. Aspects of a methodology for creating criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, April 1972. (ERIC Document Reproduction Service No. ED 099 411)

Tenopyr, M. Content-construct confusion. Personnel Psychology, 1977, 30, 47-54.

Table 1

Match of Factor Loadings With the Items Comprising
CRT Objectives for 1000 Sixth-Grade FCPS' Students

Item \ Factor	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
1	34																														
2	32																									22					
3	43																														
4	43																														
5	47																														
6	39																														
7	41																														
8	30																								37						
9	28				24																										
10	25																														
11	25																								23						
12	26																								32						
13																						23									
14		34																													
15		40																													
16		53																													
17		59																													
18		36																	21												
19		22	55	27																											
20		25	47	28																											
21			42	29															30												
22				67																											
23				60																											
24				58																											
25	21					32																									

Note: Decimal points are omitted for clarity in presentation.
Solid lines = excellent match; dotted lines = partial match; no lines = poor or no match.

Table 1 (continued)

Match of Factor Loadings With the Items Comprising
CRT Objectives for 1000 Sixth-Grade FCPS' Students

Item \ Factor	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
26																													37		
27	20				22																										
28	27																														
29					53																										
30					55																										
31						38							22																		
32	24			20	25	44																									
33							76																27								
34							32																71								
35							75																								
36							32																81								
37											30	23																			
38											33																				
39	34							26	21																	22					
40	20							21																							
41	32							32																							
42								47																							
43								50																							
44	23							40																	26						
45																	29											25			
46									24																						
47									58																						
48									56																						
49									22	39																					
50										59																					

Note: Decimal points are omitted for clarity in presentation.
Solid lines = excellent match; dotted lines = partial match; no lines = poor or no match.

Table 1 (continued)

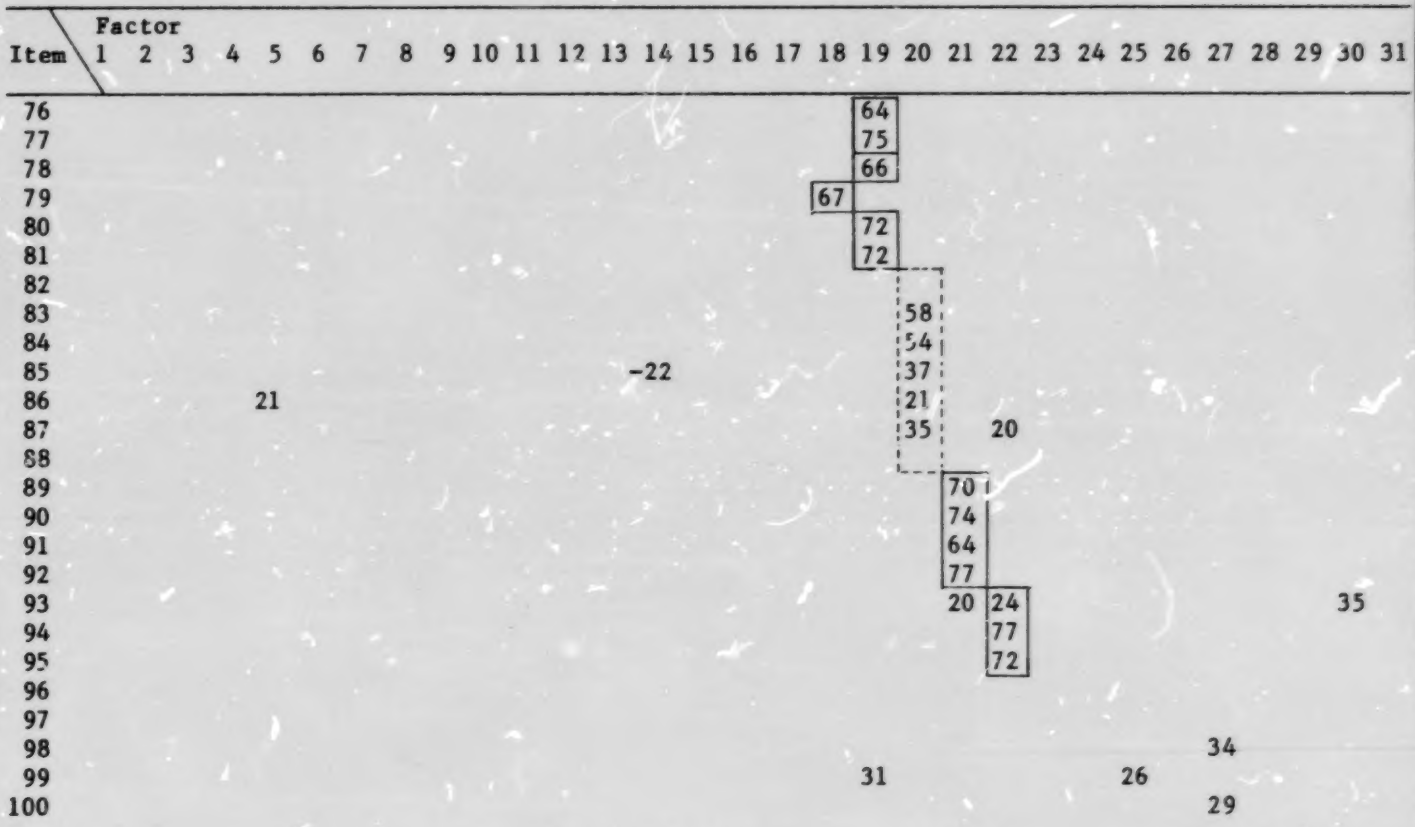
Match of Factor Loadings With the Items Comprising
CRT Objectives for 1000 Sixth-Grade FCPS' Students

Item \ Factor	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
51	25									51																					
52																								21				39			
53																								31			23				
54																								54							
55																										38					
56																										22					
57																								32							
58											47																				
59											62																				
60											49																				
61												73																			
62												73																			
63												67																			
64												20	51																		
65													65																		
66													70																		
67														-38																	
68															27																30
69															40																
70															50																
71																86															
72																85															
73																	55														
74																	64														
75																		70													

Note: decimal points are omitted for clarity in presentation.
Solid lines = excellent match; dotted lines = partial match; no lines = poor or no match.

Table 1 (continued)

Match of Factor Loadings With the Items Comprising
CRT Objectives for 1000 Sixth-Grade FCPS' Students



Note: Decimal points are omitted for clarity in presentation.
Solid lines = excellent match; dotted lines = partial match; no lines = poor or no match.

Table 2

Repeated Measures ANOVA With Pretest-Posttest Times, Classroom, and Student's Sex as Factors

sv	df*	MS	F	p<
between subjects:				
classroom	4	4127.07	14.75	.001
sex	1	345.10	1.23	ns
class x sex	4	256.35	0.92	ns
subjects w/i groups	97	279.82		
within subjects:				
times	1	19804.09	469.77	.001
class x times	4	290.44	6.89	.001
sex x times	1	18.11	0.43	ns
class x sex x times	4	73.32	1.74	ns
times x (subj w/i grps)	97	42.16		

*N = 107 sixth-grade students measured twice; thus total df = 213

Table 3

Correlation Matrix, SRA Skill Areas vs. CRT Domains
on Pretests and Posttests, for 97 Sixth-Graders

SRA Area*	FCPS' CRT Domains* 1	2	3	4	5	6	7	8	9	10	11	12	13	14
A-pre	.224	.457	.468	.271	.366	.178	.482	.463	.361	.221	.140	.493	.362	.486
A-post	.246	.444	.505	.431	.420	.362	.486	.541	.432	.405	.361	.481	.525	.438
B-pre	.113	.228	.224	.187	.095	.121	.252	.279	.259	.009	.013	.104	.270	.336
B-post	.158	.058	.108	.233	.064	.233	.095	.274	.260	.198	.296	.287	.262	.338
C-pre	.319	.306	.359	.190	.279	.171	.409	.409	.276	.206	.000	.341	.357	.437
C-post	.330	.330	.427	.439	.452	.290	.410	.441	.493	.425	.342	.304	.562	.427
D-pre	.228	.261	.309	.262	.144	.221	.365	.351	.176	.301	.128	.331	.197	.262
D-post	.070	.186	.261	.309	.082	.343	.359	.384	.347	.329	.358	.403	.341	.403
E-pre	.152	.310	.270	.138	.145	.243	.355	.371	.165	.134	.089	.321	.337	.418
E-post	.040	.000	.197	.245	.213	.216	.268	.399	.324	.258	.233	.369	.320	.284
F-pre	.328	.431	.412	.200	.352	.369	.337	.326	.357	.256	.047	.336	.303	.320
F-post	.341	.562	.492	.426	.405	.289	.479	.395	.332	.276	.238	.291	.338	.327
G-pre	.280	.515	.327	.260	.412	.164	.514	.446	.235	.162	.153	.313	.381	.263
G-post	.369	.514	.464	.465	.482	.248	.488	.469	.397	.268	.324	.421	.394	.348
H-pre	.159	.200	.524	.416	.336	.076	.324	.259	.251	.207	.153	.348	.160	.347
H-post	.224	.281	.313	.348	.258	.327	.285	.285	.392	.282	.117	.240	.230	.243
I-pre	.169	.196	.187	.227	.232	.151	.240	.160	.098	.105	.121	.332	.171	.000
I-post	.119	.267	.154	.189	.122	.188	.246	.146	.162	.116	.068	.196	.202	.199
J-pre	.286	.506	.369	.245	.424	.324	.308	.407	.179	.095	.052	.307	.264	.192
J-post	.202	.218	.273	.233	.319	.196	.305	.260	.231	.247	.164	.333	.224	.303
pretest vs posttest CRT	.268	.461	.395	.302	.313	.297	.532	.505	.397	.336	.199	.460	.531	.332